M.R.P. 1056
S.C.11/237
11th June, 1957

METEOROLOGICAL OFFICE
- 3 FEB 1958
LERWICK OBSERVATORY
SHETLAND

Forecast verification:

a critical survey of the literature

by

D. H. Johnson

(Meteorological Office)

AIR MINISTRY

METEOROLOGICAL RESEARCH COMMITTEE

At the time of issue this report has not been published. Any author desiring to quote this paper should refer to it as "A paper of the Meteorological Research Committee (London), M.R.P. No. 1056 by Mr. D.H. Johnson. A copy is available in the Library ... of the Meteorological Office."

M.R.P. 1056
S.C.II/237
11th June, 1957

AIR MINISTRY
METEOROLOGICAL RESEARCH COMMITTEE

Forecast verification: a critical survey of the literature

by
D. H. Johnson (Meteorological Office)

1. Introduction

1.1. Brier and Allen (1951) have defined forecast verification as the entire process of comparing the predicted weather with the actual weather, utilizing the data so obtained to produce one or more indexes or scores and then interpreting these scores by comparing them with some standard, depending on the purpose to be served by the verification.

1.2. Verification schemes have been devised with one or more of the following ends in view:

(i) To measure the economic value of forecasts;

(ii) To investigate the nature and cause of forecast errors and so indicate the direction in which improvement is most required;

(iii) To assess the relative merits of alternative techniques;

(iv) To provide methods of quality control by which long or short-term variations in the accuracy of forecasts can be detected;

(v) To measure the skill of individuals;

(vi) To demonstrate the value of weather services to the general public or other non-scientific bodies.

Regarding item (vi), Köppen (1893) writes: "If one will study critically the prediction service ..... A report to the public should, in this case, be as free as possible from numerical data, because figures on this subject are generally misunderstood and really carry practical convictions only in certain directions and only to specialists. For such a report, a statement as direct and as free from excuses as possible, with concrete examples and the opinion of specialists is the most convincing and suitable." Later writers endorse this view. The layman cannot be expected to understand and interpret correctly the statistical concepts appurtenant to verification; he realises that even the simplest statistical devices are open to abuse and distrusts them accordingly.

The assessment of forecasting skill (item (v)) in terms of a single numerical score with a view to classifying individuals according to their forecasting ability, has inspired a number of papers, particularly in the United States. Some writers express doubts as to the wisdom of this practice, convenient though it may be for administrative purposes. Wagemann (1935) maintained that forecasters could not be judged solely by statistics, the proper purpose of verification being not the objective assessment of skill

/but

but the detection of flaws in forecasting techniques. Petterssen (Columbia Branch, 1952) states: "Chiefs of offices are required to judge individuals but collectively we should check on our methods, not on each other ....." Schmidt (Columbia Branch, 1952) says: "..... verifications are obtained on day-by-day forecasts, not according to individuals. From these verifications come ideas for research and forecast improvement. Individual testing schemes must involve controlled experiments where everyone has the same data, for the period is never long enough to smooth out the influences of extraneous factors," and Mook (Columbia Branch, 1952): "Particularly within a given weather office, the verification system should encourage teamwork rather than pit individuals against one another."

In the present work, it has been assumed that verification procedures are required to assess the scientific and economic values of forecasts. Problems raised by items (v) and (vi) have not been primarily considered. The economic aspects are discussed in Section 2. Section 3 is concerned mainly with items (ii) to (iv).

Questions of statistical significance are not discussed in the main paper, but are dealt with in Appendix II.

2. Measuring the economic value

2.1. Perfect forecasting is unknown and the forecaster generally takes a calculated risk in issuing a particular forecast. When action is taken on the assumption that the forecast will be correct, choice of forecast for a given operation depends not only upon the meteorological probabilities but also upon the economics of the operation. The first to make this point was Peirce (1884). More recently the relevance of the economics has been discussed by Bilham (1922), Bijvoet and Bleeker (1951), Gringorten (1951), and Thompson and Brier (1955).

2.2. Gringorten (1951) gives the following approach. Suppose a forecaster must predict one of n mutually exclusive events. Let α_{ki} be the profit accruing from operations put in hand when the kth event is forecast and the ith event occurs. Let $p_1, p_2, \ldots\ldots p_n$ be the probabilities that the 1st, 2nd,.....nth events will occur, given a certain set of initial conditions. Then the profit netted, under the working assumption that event k will occur, when averaged over all forecasts based on these initial conditions, is proportional to A_k where

$$A_k = p_1\alpha_{k1} + p_2\alpha_{k2} + \ldots\ldots + p_k\alpha_{kk} + \ldots\ldots p_n\alpha_{kn}$$

There are n such relations, one for each of the n possible working assumptions. The best working assumption is the one which maximizes A_k.

For verification in purely economic terms the important factor is simply the average profit $\frac{1}{N}\sum\alpha_{ki}$. Gringorten proposes a verification score $\sum\alpha_{ki}/\sum\alpha_{ii}$, $\sum\alpha_{ii}$ being the total profit attainable through perfect forecasting.

2.3. Bijvoet and Bleeker (1951) consider in detail the case where there are only two alternatives: either to take precautions against unfavourable weather or not. Suppose during a period of N days, unfavourable weather was experienced on n days. Let the loss due to unfavourable weather on a day when no precautions were taken be L. Let the daily cost of preventative measures be P. Two cases arise depending on whether the cost of taking precautions on every day exceeds the total loss when no precautions are taken, i.e., whether $nL < \text{or} > NP$.

/Suppose

Suppose there are:

a days when unfavourable weather was forecast and did occur, cost of preventative measures = aP;

b days when the meteorologist failed to give warning of unfavourable conditions, loss = bL;

c days when unfavourable conditions were predicted and did not occur, cost of needless preventative measures = cP;

N - (a + b + c) days when no warning was given and weather was favourable, no loss, no precautions.

For $nL < NP$ the "economic benefit", B, of the forecasts is given by $nL - (aP + bL + cP)$, or, since $a + b = n$, $B/L = n - (n - b + c)\,P/L - b$.

The ratio B/L is termed the "economic benefit factor" and P/L, the "specific industry factor". Figure 1 is a graph, prepared by Bijvoet and Bleeker, showing the relation between B/L and P/L for the case $n = 10$, $N = 100$ for various values of b and c. An important consideration emerges from a comparison between curves $b = 2$, $c = 4$ and $b = 4$, $c = 2$. The total number of meteorological failures, $b + c$, is the same for both curves, but the economic values of the two sets of forecasts are equal only when $2P = L$. When $2P < L$ it is desirable that b should be minimized, but for $2P > L$ it is better if c is small.

A similar analysis can be made for the case $nL > NP$, when it is invariably preferable to have b minimized rather than c, subject to the total number of failures being constant.

2.4. The foregoing treatments illustrate the type of approach taken in verifying forecasts for individual commercial undertakings when the economic circumstances are known. Often it is impossible to determine the economic value of the forecasts. For example, one could not assess the monetary value of the gale warning service. Further, it is impracticable to verify on an economic basis when the same forecasts must serve large numbers of recipients whose interests diverge and conflict. Moreover, such scores as the average profit will not reflect the scientific worth of the forecasts. A general scheme of verification requires a system which measures simply "the approach to the truth" (Brier and Allen, 1951). Such systems are discussed in the following section.

3. Scientific value of forecasts

3.1. Verification techniques are required to meet the following circumstances:

(i) a choice is made between one of two alternatives (dichotomous classification);

(ii) a selection is made between several mutually exclusive alternatives (manifold classification);

(iii) the forecast and observed values form continuous distributions;

(iv) a composite forecast which refers to several elements is made; this is often written but sometimes contains numerical values;

(v) the future pressure pattern, together with positions of fronts, is specified on a field basis.

No single formula or procedure can cater for the diversity of circumstances listed above. The first four possibilities will be considered in turn. The fifth class is rather special and is omitted from the present survey.

/3.2.

3.2. Dichotomous classification

3.2.1. A surprisingly large range of forecasts can be adapted for verification in dichotomous (black or white) terms. Some phenomena, such as thunder, can only be treated on an occurrence or non-occurrence basis. Other elements, rain, wind and temperature for example, form continuous distributions but can nevertheless be verified by dichotomous classification: rain/no rain; gale/no gale; frost/no frost; temperature, above/below normal. Crossley (1954) used "black or white" indexes to verify four-day forecasts (change or no-change of weather type), forecasts of development of secondary depressions, warnings of ice on conductor rails, and visibility forecasts.

3.2.2. The information required in verifying black or white forecasts can be expressed in a 2 x 2 contingency table as follows:

		Forecast		
		Black	White	Total
Observed	Black	c	o - c	o
	White	p - c	s - o -p + c	s - o
	Total	p	s - p	s

This notation was given by Gilbert (1884). Other writers have used alternative layouts but Gilbert's notation will be used throughout the following discussion. The problem of deriving a satisfactory verification index is essentially that of measuring association between attributes, which is discussed from a statistical standpoint by Kendall (1948).

The forecast and observed events can be considered independent if there is the same proportion of black (or white) occurrences amongst the forecasts of black, as amongst the forecasts of white, i.e., if $c/p = (o - c)/(s - p)$ or $cs - op = o$. Taking an alternative view: o/s is the expectation that black will occur on any day of the trial period, so op/s represents the expected number of correct forecasts when p forecasts of black are issued at random. If $c > op/s$, then the observed and forecast events are positively associated; if $c < op/s$, there is negative association. The quantity $cs - op$ occurs in the numerator of several proposed indexes of forecasting success.

Verification can proceed from two complementary aspects. One can examine (i) the probability that a prediction will be fulfilled, and (ii) the chance that an occurrence will be predicted. A measure of (i) is c/p; (ii) is measured by c/o. (Doolittle, 1885; Grossmann, 1895; etc.). The Weather Division, U.S.A.A.F. (1944) has termed aspect (i), the "postagreement" and (ii), the "prefigurance". These terms are inelegant, but no others have been suggested. Which is the more important aspect depends on the use to be made of the forecasts. A film director, awaiting fine weather for a costly day on location, might not mind missing a few good days if he could be certain that when a fine day is confidently forecast, a fine day will occur. For him postagreement is essential. On the other hand, in a typhoon warning service, it is essential that all storms which occur should be forecast, even though some warnings are not fulfilled. Here prefigurance is more important than postagreement. In a general purpose index, postagreement and prefigurance should receive equal weight. The following paragraphs discuss the merit of proposed verification coefficients.

/3.2.3.

3.2.3. Percentage correct. The percentage of the forecasts which are correct is given by $100\ (2c + s - o - p)/s$. This simple score has some virtue if blacks and whites have equal probabilities of occurrence, but it can be misleading when one of the probabilities is small. An oft-quoted example concerns Finley's verification of tornado prediction (Finley, 1884). Tornadoes were forecast on 100 out of 2803 occasions. 51 storms occurred and 28 warnings were fulfilled. Finley remarks on the high percentage correct (96.6), but several writers have pointed out that the percentage correct would have been 98.2 if "no tornadoes" had been forecast perpetually. Yet the predictions clearly had some merit since $28/100 > 51/2803$ ($c/p > o/s$).

3.2.4. Gilbert's scores. Gilbert (1884) proposed two indexes. The first was the ratio of the successful forecasts of black to the sum total of the forecasts of black plus the unforecast occurrences of black: $c/(p + o - c)$. This is a bad index from several points of view: (i) a different answer is obtained if white forecasts are considered instead of black (the formula is incapable of inversion); (ii) the index varies between zero and one but the expected value for the random issue of p black forecasts is $op/(sp + so - op)$, which varies with o/s; (iii) a perpetual forecast of black gives a value o/s, which is high when $o \gg (s - o)$.

Gilbert's second index is based upon the first, the number of successful black forecasts, c, being replaced by $c - op/s$, which represents the difference between the total number of correct forecasts of black and the number expected to be correct by chance. The formula is:

$$\frac{c - op/s}{(c - op/s) + (p - c) + (o - c)} \quad \text{or} \quad \frac{cs - op}{s\,(o + p - c) - op}$$

This index overcomes the drawbacks of Gilbert's first index, but its lower limit, $-\,o\,(s - o) / (s^2 - os + o^2)$, is not fixed.

3.2.5. Doolittle's scores. Recognising the significance of c/p and c/o, Doolittle (1885) formed an index c^2/op, but this suffers from the defects of Gilbert's first score. Doolittle then argued that out of p predictions, op/s successes could be ascribed to chance, so that the quantities $p - op/s$ and $o - op/s$ represent "the fields which chance leaves for science to conquer".

The first index was therefore amended to become:

$$\frac{(c - op/s)}{(o - op/s)} \cdot \frac{(c - op/s)}{(p - op/s)} = \frac{(cs - op)^2}{op\ (s - o)\ (s - p)}.$$

The latter form has much to recommend it. It takes values between zero and one, zero being the score attained through a random issue of forecasts. A drawback is that a value of one is obtained not only for a completely correct set of forecasts, but also for complete failure. Doolittle justifies this by remarking that complete failure is as difficult to attain as complete success. While this is true, it is still undesirable to use an index which does not conform with the ordinary idea of success.

3.2.6. Peirce's index. Peirce (1884) suggested use of the expression:

$$\frac{c}{o} - \frac{(p - c)}{(s - o)} = \frac{cs - op}{o\ (s - o)}$$

c/o is the probability of a black event being correctly forecast; $(p - c)/(s - o)$ is the probability that a white event will be forecast as black. Peirce's formula avoids most of the pitfalls and varies between -1 and +1, -1 corresponding to complete failure and +1 to complete success. It can be criticised for giving too much weight to prefigurance.

/3.2.7.

3.2.7. Lacour's index. Lacour (1906) in verifying Danish storm warnings, used the ratio of the probability of a black event following a black prediction to the probability of a black event following a white prediction. Symbolically, the ratio is:

$$\frac{c}{p} \cdot \frac{(o - c) + s - (o + p - c)}{o - c} = \frac{c\,(s - p)}{p\,(o - c)}$$

This score is open to abuse by the cautious forecaster, who can obtain a high value by forecasting white only when it is almost certain to occur. If black and white events are inverted a different value is obtained.

3.2.8. Clayton's index. Clayton (1934) proposed the use of the difference between the probability of a black event following a black prediction and the probability of a black event following a white prediction, i.e.:

$$\frac{c}{p} - \frac{o - c}{(o - c) + s - (o + p - c)} = \frac{cs - op}{p\,(s - p)}$$

This score is complementary to Peirce's index in forming part of Doolittle's formula. It weights postagreement rather than prefigurance.

3.2.9. Crossley's indexes. Crossley (1952) termed c/p the "useful effort" and c/o the "forecast accuracy". He suggested that when economy of effort is the main consideration and time of lesser importance the forecasts should maximise the "useful effort". Realising that the "useful effort" was not a satisfactory verification index, Crossley (1954) proposed the quantity,

$$\frac{1}{2}\left\{ \frac{c}{o} + \frac{s - (o + p - c)}{s - o} \right\} = \frac{1}{2}\left\{ \frac{cs - op}{o(s - o)} + 1 \right\}$$

which he termed the "usefulness".

c/o is the "forecast accuracy" of the black forecasts and $(s - o - p + c)/(s - o)$ is the "forecast accuracy" of the white forecast. The formula is basically that proposed by Peirce, apart from the constants. Crossley's index ranges from 0 to 1 with a value of 1/2 as the lower limit for worthwhile forecasting.

3.2.10. The skill score. The skill score has been much used in the U.S.A. in recent years. It is usually given in the form:

$$S = \frac{\text{number of correct forecasts} - \text{number correct by chance}}{\text{total number of forecasts} - \text{number correct by chance}}$$

and is attributed to Heidke (1926). Symbolically,

$$S = \frac{c + s - o - p + c - \left[op/s + (s - p)(s - o)/s\right]}{s - \left[op/s + (s - p)(s - o)/s\right]} = o\,\frac{2\,(cs - op)}{(s - p) + p(s - o)}$$

S is zero if forecasts are no better than chance or if either black or white is perpetually forecast. S is one for perfect forecasting except in the special case when no blacks occur and none is forecast, when the score is undefined. The possible range of negative values is not fixed: the score for complete failure is $-2o\,(s - o)/(o^2 + (s - o)^2)$, which depends upon the observed proportion of blacks and attains its minimum value of -1 only when $o = s/2$.

A diagram which permits rapid calculation of skill scores has been prepared by Jorgensen (1956).

/3.2.11.

3.2.11. Improved percentage correct. Bleeker (1946) remarked that the percentage correct could be rendered less susceptible to abuse if the successful black and white predictions were weighted in proportion to the frequencies of black and white occurrences. The index thus obtained is:

$$\frac{c\,(s - o)/s + (s - o - p + c)\,o/s}{c\,(s - o)/s + (p - c) + (o - c) + (s - o - p + c)\,o/s}$$

Bleeker himself rejects the index on the grounds that the expected value for random forecasting, namely $o(s - o)/(sp + 2os - 2op + o^2)$, is not fixed.

3.2.12. Yule's indexes. The indexes hitherto discussed were devised with the meteorological problem in mind. In the early years of the present century, the task of devising measures of association between attributes drew the attention of several eminent statisticians. Yule (1912) in a comprehensive review, quotes the following measures of association:

(a) The coefficient of association $Q = \dfrac{c(s - o - p + c) - (o - c)(p - c)}{c(s - o - p + c) + (o - c)(p - c)}$

$$= \frac{cs - op}{c(s - o - p + c) + (o - c)(p - c)};$$

(b) The coefficient of colligation $Y = \dfrac{\sqrt{c\,(s - o - p + c)} - \sqrt{(o - c)(p - c)}}{\sqrt{c\,(s - o - p + c)} + \sqrt{(o - c)(p - c)}};$

(c) $$V = \frac{c\,(s - o - p + c) - (o - c)(p - c)}{\sqrt{op\,(s - o)\,(s - p)}}$$

$$= \frac{cs - op}{\sqrt{op\,(s - o)(s - p)}}\ .$$

These three coefficients vanish when the attributes are independent ($cs = op$), take values ranging between +1 and -1, and decrease as the cell frequencies proceed from association to disassociation. Q and Y become +1 when $p = c$ or $c = o$, and -1 when either c is zero or $(s - o - p + c)$ is zero. V becomes +1 only when $p = c$ and $c = o$, and -1 only when c is zero and $(s - o - p + c)$ is zero.

Kendall (1948) defines the condition in which all As are B, though not all Bs are A as "complete association". The state when all As are B and all Bs are A is termed "absolute association". Thus Q and Y are +1 for both complete and absolute association. V becomes +1 only for absolute association. On account of this difference Kendall prefers Q and Y to V. The preference is understandable in the common circumstances when the observer or experimenter has no control over the relation between the attributes. In meteorology however, one attribute is determined by the forecaster and it may not be difficult for him to "play the system" so as to obtain complete association even though the majority of his forecasts are valueless. For general verification it is desirable that an index should achieve its highest value for absolute association alone, and V is therefore preferred to Q or Y.

To derive V, Yule used the concept of correlation. The attributes can be regarded as variates, which take only the values zero and one. V then represents the coefficient of correlation between the two variates. V can be interpreted in terms other than those of correlation since it is also given by the square root of Doolittle's index or by the square roots of Peirce's and Clayton's indexes in combination. It gives equal weights to postagreement and prefigurance and is closely related to the important chi-squared statistic: for a 2 x 2 table, $\chi^2 = sV^2$. Chi-squared itself is a measure of the strength of the relation between attributes but since chi-squared is a function of the total number of observations as well as of the relative proportions of the cell frequencies, it is strictly a measure of statistical significance* and is not suitable for use as a coefficient of association.

*See Appendix II

/In

In common with the skill score and other indexes, V is undefined if black (or white) perpetually occurs and is perpetually forecast. In such circumstances a measure of association is hardly called for. V is always numerically greater than or equal to the skill score.

3.2.13. Of the indexes discussed V is perhaps the most suitable for use as a general verification index, both in view of its intrinsic properties and because it is closely linked with well-understood statistical quantities. The use of V in forecast verification was advocated by Wallén (1921) and later by Bleeker (1946). Bleeker attributes the formula to Wallén, although Wallén himself acknowledges the debt to Yule. Heidke (1928) criticised Wallén's formula (Yule's V) on the grounds that it was not based on probability calculations: it did not explicitly measure the improvement over chance as did the skill score. The criticism is hardly justified. While it is not obvious from the correlation derivation of V, it is clear from the alternative derivations following Doolittle or Peirce and Clayton, that chance successes are allowed for implicitly.

Values of Peirce's index, Clayton's index, Yule's V and the skill score for some sample contingency tables are given below. The effects of the relative weightings given to prefigurance and postagreement are apparent.

1)

		Forecast B	Forecast W	Total
Observed	B	10	0	10
	W	10	80	90
	Total	20	80	100

2)

		Forecast B	Forecast W	Total
Observed	B	10	0	10
	W	60	30	90
	Total	70	30	100

3)

		Forecast B	Forecast W	Total
Observed	B	10	0	10
	W	80	10	90
	Total	90	10	100

4)

		Forecast B	Forecast W	Total
Observed	B	1	9	10
	W	0	90	90
	Total	1	99	100

5)

		Forecast B	Forecast W	Total
Observed	B	1	0	1
	W	1	98	99
	Total	2	98	100

Contingency Table No.	Peirce's Index (prefigurance)	Clayton's index (postagreement)	Yule's V	Skill Score
(1)	.89	.50	.67	.62
(2)	.33	.14	.22	.09
(3)	.11	.11	.11	.02
(4)	.10	.91	.30	.08
(5)	.99	.50	.70	.66

/3.2.14.

3.2.14. Comparative indexes. Some authors (e.g., Clayton, 1889; Weather Division, H.Q., U.S.A.A.F., 1944) have objected to the calculation of the chance successes from the observed frequencies for the trial period. They maintain that the climatic normals should be used as the standard of comparison, since the use of observed frequencies implies a degree of foreknowledge of the observed weather. It is difficult to see the logic of this argument. It is true that before the event, an estimate of the expected success of forecasts issued at random is given by the climatic normals. Nevertheless, in assessing, post facto, the number of forecasts which could have been correct by chance during the trial period, the determination of the expected value must be based on the observed frequencies. The chance forecasts are assumed by definition to have been made without foreknowledge.

It is often desirable to compare the success of a forecasting system with the success of less subtle (and usually less costly) techniques such as predicting persistence or the climatic normals. Some writers have devised indexes which measure the success relative to the simpler technique. Heidke (1926) proposed the use of an index $G = (E - B)/(1 - B)$, where E is the number of correct forecasts expressed as a fraction of the total number of forecasts, and B is the number of correct persistence forecasts also expressed as a fraction of the total number of forecasts. B could be replaced by any convenient standard. If the fraction correct by chance is substituted for B the skill score is obtained. Heidke (e.g., 1935) repeatedly stressed his preference for a persistence base because persistence normally gives a higher score than pure chance or climatology.

Crossley (1954) suggests that improvement over persistence, or any other forecasting system, be measured by the index $J = (I - I_o)/(1 - I_o)$ where I is the value, appropriate to the system under test, of some verification index taking a maximum value of one, and I_o is the value scored by the standard.

A distinction should be recognised between indexes discussed in previous sections which measure the association between observed and forecast events, and indexes such as Heidke's G or Crossley's J which measure the relative success of forecast systems. Whether indexes of the latter type have much value is a moot point. Statements of the verification statistics for the forecast technique and for persistence, separately, contain more information and are better interpreted than hybrid indexes which measure relative success.

The accuracy of conventional forecast techniques is higher during periods when the weather displays a greater degree of persistence. Indexes which measure gain over persistence have therefore been used in attempts to make comparisons between verification data obtained for different periods more fair. There are two objections to this practice. Firstly, it may be found that the greater the variability of the weather, the more the forecaster gains relative to persistence (cf. remarks made in section 3.4.8. regarding Priestley's index). Secondly, it cannot be admitted that the inherent forecasting difficulty posed by a synoptic sequence is to be measured simply in terms of persistence. A suggestion due to Brier and Allen (1951), that indexes can be adjusted for difficulty using a regression equation $I' = I + a\,(Ip - \overline{Ip})$, where Ip is the persistence score corresponding to a forecasters score I, is open to similar objections. It is preferable to avoid these difficulties by verifying alternative techniques on the same series of weather situations. If the problem must be met, for example, if the variation of forecast accuracy with time is to be measured, periods compared must be long enough to contain well-varied selections of weather types and allowance must be made for diurnal and seasonal effects.

For historical interest mention is made of two other types of standard with which routine forecasts have been compared. Mascart (1922) had forecasts made for each day of a year, in advance, a knowledge of local climate being assumed. Dinies (1936) used daily forecasts made in 1929 and 1933 as forecasts for 1934.

/3.3.

3.3. Manifold classification

3.3.1. Categorical forecasts

3.3.1.1. The data essential to the verification of forecasts which represent a choice made from several mutually exclusive alternatives can be expressed in a p x p contingency table:

Forecast Class

Observed class	A_1	A_2	...	A_i	...	A_p	Total
B_1	X_{11}	X_{21}	...	X_{i1}	...	X_{p1}	(B_1)
B_2	X_{12}						(B_2)
⋮	⋮			⋮		⋮	⋮
B_j	X_{1j}	...		X_{ij}	...	X_{pj}	(B_j)
⋮	⋮			⋮		⋮	⋮
B_p	X_{1p}	...		X_{ip}	...	X_{pp}	(B_p)
Total	(A_1)	(A_2)	...	(A_i)	...	(A_p)	N

X_{ij} is the number of occasions when class j occurred, class i having been forecast. (A_i), (B_j) are the forecast and observed frequencies of classes i and j respectively. N is the total number of occasions.

Classes A_i and B_j are defined as independent when

$$\frac{X_{ij}}{(A_i)} = \frac{(B_j)}{N}$$

For complete success entries should occur only in the leading diagonal of the table. The percentage correct is given by the ratio of the sum of the leading diagonal entries to the total number of occasions. Reasons were given in para. 3.2.3. for interpreting this statistic with caution. When the classes are formed by division of a range of values of some variate, the number falling into each class depends on the number of classes formed. The percentage correct can be varied by adjusting the number of classes, although the basic data remain unchanged.

The contingency table was one of the earliest tools employed in verification. The Deutsche Seewarte introduced tabular summaries as a routine (Köppen, 1884). Clayton (1889) used them to show where the greatest need for improvement lay. With a similar end in view, Ekholm (1907), in verifying storm warnings, computed values of $X_{ii}/(A_i)$ and $X_{ii}/(B_i)$ for a series of values of the observed Beaufort force. $X_{ii}/(A_i)$ and $X_{ii}/(B_i)$ measure postagreement and prefigurance respectively.

Bleeker (1946) points out that "black or white" verification indexes can be computed for each class, black forecasts corresponding to forecasts of the class in question and white forecasts, to forecasts of all other classes. By this method, which could even be applied when the forecast classes overlap, the variation of success between the classes can be examined. Crossley (1954) used a similar device to verify forecasts of visibility at Northolt Airport. The visibility forecasts were classified dichotomously, taking a succession of visibility limits as the dividing lines. Crossley's index (para. 3.2.9.) was computed for each limit.

/3.3.1.2.

3.3.1.2. The skill score. Most of the measures of association discussed in section 3.2. apply only to black or white forecasts and cannot be generalised simply, to cover the manifold case. One exception is the skill score. The formula,

$$S = \frac{(\text{No. of correct forecasts}) - (\text{No. correct by chance})}{(\text{Total no. of forecasts}) - (\text{No. correct by chance})}$$

becomes for manifold classification,

$$S = \frac{\sum X_{ii} - \sum (A_i)(B_i)/N}{N - \sum (A_i)(B_i)/N}$$

S varies from +1 for complete success to a negative score ($\geqslant -1$), which depends upon the observed frequencies, for complete failure.

3.3.1.3. The skill index. The Weather Division, H.Q. U.S.A.A.F. (1944) refer to a "skill index" given by,

$$I = \left[\sum_i X_{ii}\{N-(B_i)\} + \sum_i X_{ii} - N\right] \Big/ N(p-1)$$

I is +1 for complete success, zero if the diagonal frequencies take their "chance" values, and $-1/(p-1)$ when the diagonal frequencies are zero.

When applied to the "black or white" case, $p = 2$ and the formula reduces to $(cs - op)/o(s - o)$ in the notation of section 3.2. This is Peirce's index (para. 3.2.6.).

3.3.1.4. The skill score and the skill index depend only upon the frequencies in the leading diagonal and the border frequencies of the contingency table. They are applicable to the special case (statistically speaking) of the bivariate frequency table: it is assumed that the classes are arranged in order, being formed by division of the range of some variate. No allowance is made for "near misses"; the indexes depend upon the hits alone and do not explicitly measure the extent to which the shots are grouped around the target. The values of the indexes depend upon the number of classes into which the data are divided.

3.3.1.5. Vernon's deviation and quadratic skill scores. Vernon (1953) devised two scores which weight the incorrect forecasts according to the degree of error. The deviation skill score is defined as,

$$S_d = \frac{\sum_{ij} (A_i)(B_j) d_{ij}/N - \sum_{ij} X_{ij} d_{ij}}{\sum_{ij} (A_i)(B_j) d_{ij}/N}$$

where d_{ij} is equal to the number of class intervals by which the cell in the ith row and jth column is removed from the "perfect hit" cell in the same column. $d_{ij} = |j - i|$. The $(A_i)(B_j)/N$ are the expected frequencies.

/The

The quadratic skill score is

$$S_q = \frac{\sum_{ij}(A_i)(B_j)d_{ij}^2/N - \sum_{ij} X_{ij} d_{ij}^2}{\sum_{ij}(A_i)(B_j)d_{ij}^2/N}$$

The two scores are analogous to the conventional skill score, to which they reduce in the "black or white" case. They differ from each other only in degree of weighting, which is, of course, quite arbitrary. The scores become +1 for complete success and are zero when success is no better than chance. Table 1 contains a hypothetical example, taken from Vernon's paper, of the application of the scores to two contingency tables.

A common fault of systems which award marks for near misses, is that they encourage the forecaster to play safe. Professor J.G. Bryan, who collaborated in U.S. Air Force experiments in verification, has said that when attempts were made to allot partial credit to a nearly correct forecast, the scheme was compromised to a point where it was possible to "play the system", usually by forecasting the most frequent event or by forecasting persistence (Gringorten, 1951). However, Vernon's scores do not appear to suffer from this drawback in practice and Vernon shows that, for the data considered in Table 1, the forecasters would have lost by always forecasting the middle of the range (see Table 2).

3.3.1.6. Pearson's coefficient of contingency. The value of the chi-squared statistic for a p x p contingency table is

$$\sum_{ij}\left[\left\{X_{ij} - (A_i)(B_j)/N\right\}^2 \Big/ \left\{(A_i)(B_j)/N\right\}\right]$$

Chi-squared may be used to test the significance of the association displayed in the contingency table, but for reasons given in para. 3.2.12. it cannot be used as a coefficient of association. Pearson proposed a measure based on chi-squared namely,

$$C = \sqrt{\frac{\chi^2}{N + \chi^2}} \qquad \text{(Kendall 1948)}$$

C vanishes when there is no association but takes a value $((p - 1)/p)^{\frac{1}{2}}$ which depends upon the number of classes, for absolute association. This is a disadvantage from a verification point of view. It is possible to "correct" C, so that it shall have an upper limit of +1, using a scaling factor appropriate to the number of classes (Brooks and Carruthers, 1953).

3.3.1.7. Tchuprow's coefficient. Tchuprow devised a coefficient of association which like Pearson's C is a function of chi-squared, but which has the advantage of a fixed upper limit. For a p x p contingency table, Tchuprow's coefficient is,

$$T = \sqrt{\frac{\chi^2}{N(p-1)}} \qquad \text{(Kendall 1948)}$$

T is +1 for absolute association and is zero if there is no relation. For the "black or white" case T reduces to Yule's V (para. 3.2.12.).

3.3.1.8. The coefficients of Pearson and Tchuprow take all the cells of the contingency table into account, but they do not depend upon any ordering of the classes, i.e., any two rows or columns may be interchanged without affecting the value of the index. They measure association, but they do not require that a given row should be associated with any particular column. Where the classes represent divisions of a visibility or temperature scale, for example, the forecasts may have an appreciable bias and yet achieve a high coefficient of association. The indexes need not necessarily be rejected on this account; one does not reject the correlation coefficient because it is insensitive to errors of bias and scale; but the limitations must be borne in mind. A further difficulty arises because, like Doolittle's index (para. 3.2.5.), the coefficients take their minimum values for chance success and their maximum values for both complete success and complete failure. Chi-squared is undefined if a row or column total is zero.

3.3.1.9. Correlation coefficient. When the contingency table is a bivariate frequency table and the number of classes is sufficiently large, a coefficient of correlation may be calculated, due allowance being made for grouping (vide para. 3.4.4.). The correlation coefficient is sensitive to the ordering of the classes and takes negative values for negative association.

3.3.1.10. Gringorten's verification scheme. Gringorten (1951) has devised a system which forces the forecaster to select, not the most probable event, but the event whose probability is most increased in proportion to the climatic frequency. If α_{ki} is the mark awarded for predicting the kth event when the ith event occurs, then $\alpha_{ki} = 0$ for $k \neq i$ and $\alpha_{ii} = 1/Q_k$ where Q_k is the climatological probability that the kth event will occur.

The score for completely erroneous forecasting is zero. The expected value of the score for perpetual forecasting of the kth event is $\sum_{r=1}^{N} Q_r \alpha_{kr} = 1$. Similarly the score for forecasts issued at random is one. Gringorten suggests that to eliminate any gain due to forecasting persistence the "climatic frequencies" should be calculated for each basic class of antecedent weather, so that the expected score for "blind" forecasting will be unity, regardless of whether the blind forecast is a perpetual forecast, a persistence forecast or a forecast of pure chance.

The expected score for forecasting the kth event on any occasion is $\sum q_r \alpha_{kr} = q_k / Q_k$ where q_r is the true probability that event r will occur on that occasion. The forecaster is therefore encouraged to predict the event whose probability is most increased in relation to the climatological probability.

An index which is zero for "blind" forecasting and unity for perfect forecasting is given by

$$K = \frac{\sum^{N} \alpha_{ki} - N}{\sum^{N} \alpha_{ii} - N},$$

where $\sum^{N} \alpha_{ki}$ is the forecaster's accumulated score and $\sum^{N} \alpha_{ii}$ is the accumulated score for perfect forecasting.

/When

When the forecaster is permitted to predict that the event will fall into one of two or more alternative classes (i.e., he is not required necessarily to choose a single class) then a similar verification scheme can be applied. The score for correctly predicting that the event would fall into one of the classes u, v,.....z becomes

$$1 \Big/ \sum_{r=u}^{z} Q_r$$

This method of verification is unsuitable for checking routine forecasts, since the forecast selected with a view to maximising the score might often differ from that normally issued. The scheme was devised with the object of assessing the forecaster's skill and requires that a special set of forecasts be made. It might be used to test the insight into the meteorological probabilities provided by alternative forecast techniques, particularly if those techniques yielded objective estimates of the probabilities.

The opinion that certain scores leave too much room for hedging have led some (e.g., McKillip, Columbia Branch, 1952) to suggest that the forecaster should be unaware of the nature and existence of the checking scheme. By contrast, Gringorten's approach to verification, exemplified by the system presently discussed and by that given in para. 2.2., requires the forecaster deliberately to play the system. "If the scoring system influences the forecaster adversely, then it is the fault of that particular scoring system. A good system when 'played' by the forecaster will force him to use his meteorological know-how" (Columbia Branch, 1952).

Gringorten, Lund and Miller (1955) discuss the practical application of Gringorten's system to verification made over a period of two years at four U.S. Air Force stations.

Verification methods which apply solely to probability forecasts are discussed in the following section.

3.3.2. Probability forecasts

3.3.2.1. Checking systems for probability forecasts have not been so widely studied as those applicable to categorical predictions, presumably because numerical estimates of likelihood do not usually accompany routine forecasts, although subjective verbal qualifications are common. Cooke (1906) is generally credited with the first attempt to verify probability statements. The proportions of erroneous forecasts were simply compared with the degree of confidence indicated.

Brier (1944) offers a convenient form of tabulation, used in verifying 6-hr. precipitation forecasts for Washington as follows:

/Verification

Verification of statements of probability of precipitation

F/c probability of precipitation	No. of cases (n)	Average probability (p)	No. of precipitation cases: expected (np)	observed	Observed proportion of precipitation cases
.00 - .09	56	.036	2.0	1	.018
.10 - .19	38	.162	6.2	13	.342
.20 - .29	6	.222	1.3	1	.167
.30 - .39	10	.326	3.3	3	.300
.40 - .49	25	.439	11.0	13	.520
.50 - .59	9	.558	5.0	2	.222
.60 - .69	9	.654	5.9	5	.556
.70 - .79	13	.755	9.8	10	.769
.80 - .89	10	.835	8.4	8	.800

A table of this type will indicate:

(i) the correlation between forecast and observed probabilities;

(ii) errors of scale; the forecaster may be either optimistic or pessimistic in the assessment of risks;

(iii) what proportion of the forecasts fall into high or low probability classes; these are generally the most useful forecasts.

3.3.2.2. Brier's index. Brier (1950) has devised an index P for use when an event can occur in one of r possible classes,

$$P = \frac{1}{n} \sum_{j=1}^{r} \sum_{i=1}^{n} (f_{ij} - E_{ij})^2$$

where f_{ij} is the forecast probability that the event will fall in the jth class on the ith occasion. E_{ij} is an indicator which takes the value one, if the event does occur in the jth class on the ith occasion, but is zero if the event falls in some other class. Summations are taken over r classes and n events. $\sum_{j=1}^{r} f_{ij} = 1 \quad , \quad i = 1, 2, \ldots\ldots n .$

P ranges from zero for complete success to two for complete failure. Complete success requires a forecast probability of one to be correctly given on every occasion. P depends on the frequency distribution of the forecast probability statements and on the correlation between forecast and observed probabilities. If p_j is the relative frequency with which jth class occurred during the forecast trial, then the minimum score which can be got by predicting the same set of probabilities for every occasion is

$$P = 1 - \sum_{j=1}^{r} p_j^2 ,$$

which occurs when $f_{ij} = p_j$ for all i. If the jth event is categorically forecast on every occasion $P = 2 (1 - p_j)$ and if categorical forecasts are made at random, the expected value of P is $2 (r - 1)/r$. The score must be

/interpreted

interpreted with caution, since repeated issue of the climatic probabilities will give a relatively small value in some circumstances. A perpetual forecast of an event which has a probability of occurrence of .9 yields a score as low as .2.

To achieve a score S* which increases with increasing success Leight (1953) has transformed S by the relation $S^* = 1 - S/2$.

3.3.2.3. Information theory. A novel way of verifying probability forecasts, based on information theory, is given by Holloway and Woodbury (1955). In a situation with several outcomes, the "information" is defined as $I = \sum_i p_i \log_e p_i$ where p_i is the probability of the ith possible outcome. Maximum information is available when $p_r = 1$ and $p_i = 0$ for $i \neq r$; then, $I = 0$. Smaller amounts of information correspond to negative values of I. If there are n outcomes, it can be shown that I takes its minimum value when $p_1 = p_2 = \cdots\cdots\cdots = p_n = 1/n$.

Holloway and Woodbury define the Information Ratio for probability forecasts by the relation

$$(I\,R)_p = \frac{\sum_{i=1}^{n} \log_e p'_i/p_i}{\sum_{i=1}^{n} \log_e 1/p_i}$$

where p'_i is the probability forecast for the event which occurred and p_i is the corresponding climatic probability. Summations are taken over n forecasts. $(I.R.)_p$ is described as the ratio of the increase in information given by the forecast beyond that provided by climatology, to the amount of information beyond climatology required for a perfect forecast. This description does not exactly agree with the definition of "information" previously given. $(I.R.)_p$ is one for perfect forecasting and zero if the climatic probabilities are predicted but, as defined above, it becomes minus infinity if zero probability is forecast for the event which occurs.

Information theory can also be applied to the verification of categorical forecasts. The Information Ratio is then defined by the relations

$$(I.R)_c = \frac{\sum_i \sum_j \frac{X_{ij}}{N} \log_e \frac{X_{ij}}{(A_i)} - \sum_j \frac{(B_j)}{N} \log_e \frac{(B_j)}{N}}{-\sum_j \frac{(B_j)}{N} \log_e \frac{(B_j)}{N}}$$

$$= \frac{\sum_i \sum_j X_{ij} \log_e X_{ij} - \sum_i (A_i) \log_e (A_i) - \sum_j (B_j) \log_e (B_j) + N \log_e N}{N \log_e N - \sum_j (B_j) \log_e (B_j)}$$

using the notation of Section 3.3.1.

/Some

Some of the properties of $(I.R.)_c$ differ from those of $(I.R.)_p$. $(I.R.)_c$ is one for perfect forecasting but zero is the value expected for forecasts issued at random. Using the categorical forecast definition it is possible to attain the maximum score for forecasts which are completely wrong since $(I.R.)_c$ is a measure of association which is not concerned with the ordering of the classes (cf. Pearson's C and Tchuprow's coefficient, Section 3.3.1.).

Holloway and Woodbury compare the Information Ratio with the skill score by computing the indexes for sample tetrachoric tables. Skill scores of zero and one correspond to Information Ratios of zero and one, but an Information Ratio of 0.5 corresponds to skill scores ranging from .71 to .78. The authors state: this implies one can make a very good forecast having only half the information required for a perfect forecast. This questionable remark illustrates the danger of attaching a plain language label which has a broad everyday connotation to a verification score; the Information Ratio is a mathematical function only loosely connected with the subjective notion of information; it could be variously defined so as to give several different answers for the fraction of information available. Heidke's "value" (para. 3.4.6.) and Crossley's "usefulness" (para. 3.2.9.) are other examples of names which may mislead.

3.4. Continuous distribution

3.4.1. If the forecast is expressed as a single value chosen from a continuous distribution or a variate scale, the simplest verification statistic is the forecast error. The distribution of forecast errors is examined using some sort of frequency curve, either a histogram or a cumulative frequency curve or both. Alternatively, the information can be given in tabular form. Important features of the error distribution are measured by the mean forecast error or "bias" and the root-mean-square (r.m.s.) error. These statistics provide convenient methods of discriminating between error distributions of alternative techniques; their standard errors can readily be estimated, even when the errors themselves are serially correlated.

3.4.2. R.m.s. error. The r.m.s. error of a series of forecast errors $x_1, x_2 \ldots\ldots\ldots\ldots x_n$ is $\left[\frac{1}{N}\sum_{r=1}^{N} x_r^2\right]^{\frac{1}{2}}$. This can be expanded as $\left[\frac{1}{N}\sum_{r=1}^{N}(x_r-\bar{x})^2 + \bar{x}^2\right]^{\frac{1}{2}}$ i.e., the r.m.s. error contains contributions from the bias and the standard deviation of the forecast errors. The bias is usually small, so the r.m.s. error is largely a measure of dispersion. An alternative measure of dispersion is the mean modulus of the errors, but the r.m.s. error gives more weight to the larger errors and is more convenient for statistical manipulation.

It is sometimes objected that the r.m.s. error encourages the forecaster to play the middle of the range. An oft-quoted example concerns the forecaster faced with the problem of predicting the night minimum temperature. He may be reasonably sure that the minimum will be 30 degrees if cloud disperses and 40 degrees if cloud remains. To play safe he forecasts 35 degrees even though such a minimum is unlikely. In some circumstances, for example in forecasting route winds for aviation, where safety considerations are involved, it may be desirable that the forecaster should avoid the risk of making large errors by cutting down on the bad failures at the expense of the spectacular successes.

3.4.3. Correlation coefficient. The correlation coefficient is the measure of association most commonly used in measuring association between variates. It is insensitive to errors of bias and scale and its function is complementary to those of the bias and r.m.s. error.

/If

If $F_1, F_2 \ldots F_N$ and $O_1, O_2, \ldots O_N$ are series of forecast and observed values, the coefficient of correlation between them is

$$\frac{\sum_{r=1}^{N}(F_r-\bar{F})(O_r-\bar{O})}{\sqrt{\sum_{r=1}^{N}(F_r-\bar{F})^2 \sum_{r=1}^{N}(O_r-\bar{O})^2}}$$

When F and O are vectors this formula gives the **stretch** vector correlation coefficient (Durst, 1954), provided that the product terms are scalar products. Both the correlation coefficient and the **stretch** vector correlation coefficient take values between -1 and +1, being zero when there is complete linear independence; but for the **stretch** vector correlation to be maximised, corresponding pairs of vectors must have the same direction as well as satisfy a linear relation.

Correlations between forecast and observed values contain components due to climatic and synoptic trends which are common to both sets of data. A more sensitive coefficient is got when the forecast and observed changes are compared. Complications due to trends are then less intrusive, unless there is a marked diurnal variation which has not been eliminated by choice of forecast period.

3.4.4. The bias, r.m.s. error and correlation coefficients are numerical digests from which characteristic properties of the error distribution can be inferred. For some purposes it will be important to know the percentage of errors which fall within or beyond specific limits, in which case reference should be made to the basic frequency distribution of errors. It may also be pertinent to examine the two aspects, postagreement and prefigurance discussed in previous sections. The Weather Division, U.S.A.A.F. (1944) observe that the distribution of postagreement and prefigurance are cross-sections of the bi-variate surface of forecast and observed values. The two distributions are not generally the same, nor will either distribution remain invariant over the whole range of events.

In some cases the importance of an error of given size varies with the range of the element in question. Murray and Smith (1956) remark that an error of a few degrees may be vital when temperatures are around freezing, but is relatively unimportant on a hot summer's day. Likewise Best and Panofsky (1956) on forecasting precipitation amounts: an error of half an inch of precipitation is much more important if the precipitation is observed to be zero than if it is observed to be five inches; for such types of forecast, it is better to define a number of categories for the forecast variable, which cover larger ranges for larger precipitation amounts, and find the actual number of categories by which the prediction missed the actual amount. Similar considerations lead Brier and Allen (1951) to mention the possibility of expressing errors in such elements as cloud height and visibility as percentages of the actual value.

3.4.5. Comparative indexes. Comments on the use of special indexes in assessing the relative success of forecast techniques were made in para. 3.2.14. Similar indexes which apply particularly to the type of forecast discussed in the present section have been devised by Heidke (1929) and Priestley (1945).

/3.4.6.

3.4.6. Heidke's index. Heidke's index, W, which he terms the "value" is defined as

$$W = \frac{\sum|x'| - \sum|x|}{\sum|x|} ,$$

where x' is the forecast error and x is the error given by persistence.

3.4.7. Priestley's index. Priestley defined a "percentage success", P, by the relation

$$P = 100\left[1 - \frac{\epsilon^2}{\sigma_t^2}\right]^{\frac{1}{2}} ,$$

where ϵ is the r.m.s. error of the forecasts and σ_t is the r.m.s. error of corresponding persistence forecasts.

Sawyer (1953) remarks that this index is particularly flattering to the forecaster because a small reduction from one in the ratio ϵ/σ_t produces an apparently large success; for example if $\epsilon = .8\,\sigma_t$, P = 60%.

It should not be forgotten that the percentage success measured by Priestley's P is purely relative. For example, the "percentage successes" of routine 18-hr. wind forecasts at 30,000 ft. and 50,000 ft. are 75 and zero respectively; yet the forecast errors are actually less at 50,000 ft. than at any lower level. It was mentioned in para. 3.2.14. that the errors of most techniques vary from period to period, being greater at times when persistence errors are large. For this reason, indexes such as Priestley's are sometimes used to make tests of forecasts issued during different periods or seasons more comparable. However, Murray (1954) has indicated that the "percentage success", when computed for upper wind forecasts, is not independent of the winds' variability, high values of P being got when persistence errors are large. Murray concludes that large changes can usually be forecast better than small changes, although the absolute errors are greater for the large changes.

3.4.8. For some purposes it may be convenient to classify data which form a continuous distribution and treat them by the methods of Section 3.3. A system developed by the Weather Information Branch, H.Q., U.S.A.A.F. (1943) is an example of this type of approach. The main purpose of the system was to classify personnel but it contains features of more general interest. The first stage was to divide the distribution of an element into thirty intervals of equal probability (trentiles) on the basis of past records. No complications arise for an element such as atmospheric pressure which, when measured to the nearest tenth of a millibar, takes very many values over a wide range. In less normal distributions a single value may occupy a number of trentiles. For example, the authors give a rainfall case where the "no rain" category occupies twenty trentiles; "trace" requires three trentiles, whilst all other possible rainfall amounts are contained in the remaining seven trentiles. The mark awarded for each forecast is simply the number of trentiles which separate the forecast and observed values. When a value which occupies one trentile or less is correctly forecast, the mark is zero. When a value which occupies more than one trentile is correctly forecast, the mark awarded is $(n2 - 1)/30$, where n is

/the

the number of trentiles occupied. $(n^2 - 1)/3n$, the "deviational penalty score", is the mark which would be got by random forecasts within a group of n trentiles. If the forecast value differs from the observed value, and either the forecast value or the observed value or both occupy more than one trentile, then the mark awarded is $\frac{1}{2}(n_1 + n_2) + K$, where n_1 and n_2 are the number of trentiles occupied by forecast and observations and K is the number of trentiles separating those containing the two values. This system of penalties compensates for the relatively greater ease with which some categories can be predicted. Taking the rainfall data mentioned above, a correct forecast of "no rain" receives a high penalty mark (6.7) because two out of three "no rain" predictions will be correct by chance. A correct forecast of "trace" has a low penalty mark (0.9) because only one in ten will succeed by chance.

When the above scheme had been in operation for some time, it was found that the average scores achieved varied with the difficulty of the weather situation (Weather Inf. Branch, H.Q. U.S.A.A.F., 1944). Each individual score was therefore standardised by subtracting from it the mean of all the scores for that day and dividing the difference by the standard deviation. Standardisation facilitates combination of scores and adjusts for difficulty, but standardised scores are useless for measuring improvements due to changes of technique, etc.

3.4.9. To afford a comparison between some of the verification indexes discussed in sections 3.2., 3.3. and 3.4., values have been computed for a series of 24-hr. temperature forecasts. These are examined in Appendix I.

3.5. Written and composite forecasts

3.5.1. Most weather forecasts specify the future state of several elements. A number of verification systems have been designed to give a single index of the success of the whole forecast. In many cases the predictions are not numerical but are expressed in words. Inevitably these systems are more arbitrary than those so far discussed.

3.5.2. The verification rules for an early scheme used by the U.S. Signals Office are listed by Harrington (1899). A verifying officer estimated the areas of a forecast region affected by rain, warmer or colder conditions, etc. consulting synoptic charts for two hours of the 24-hr. forecast period. On this basis, the percentage success was estimated subjectively for each of the factors: weather and temperature; display of cautionary and storm signals; warnings of cold waves. Weights of five, two and one respectively were given to these factors in computing "the general monthly percentage of verifications of all indications". The use of subjective estimates was evidently unsatisfactory, for Harrington comments: ..... the low average of verification of late years as compared with early years of the service is due to increased care and precision in estimating the verifications.

Another subjective check which has been used from time to time in various meteorological services is known as the jury method. Verification is made by a small committee who award the forecast some mark on an arbitrary scale, perhaps ranging from zero for failure to five or ten for close agreement. Resort to the jury method is made when the meteorologists have difficulty in laying down acceptable objective criteria, but subjectivity does not solve the problems, although it may obscure them. Fresh difficulties are raised; a subjective mark does not carry with it the climate of thought which led to its fixing and there is no knowing afterwards, or perhaps even at the time, just how much weight was given to the different factors. Variations with time of subjectively fixed indexes may be due to changes in forecast success or to changes in the critical standards of the assessors. An objective system may have its flaws but they will bear investigation and its judgment will not lack in consistency. Freybe (1911) defended subjective verification on the grounds that it was better than no verification at all.

/3.5.3.

3.5.3. The desire to eliminate subjectivity led to two developments. Firstly, some weather services prescribed all the terms and phrases which were to be used in forecasting, each being given a specific meaning. Secondly, verification systems increased in complexity. Milhan (1942) quotes the current U.S. Weather Bureau verification regulations. For precipitation these were:

"Rain or snow forecast and all modifications thereof containing the terms 'possibly', 'probably', etc. will be verified by the occurrence of .01" or more of precipitation. Precipitation to the amount of .01" or more which is not forecast or which is forecast and does not occur will be charged as a failure. A forecast of precipitation will be considered as neither a success nor a failure when a trace of precipitation is recorded within the period specified."

In determining the percentage of verification, the number of forecasts verified will constitute the dividend, the divisor will be the sum of the forecasts verified, the number of forecasts not verified and the number of occurrences of .01" or more of precipitation which were not forecast. Corresponding rules were given for temperature, storm warnings, etc. and the individual percentages were then averaged to give an overall figure. On the basis of this system, the Weather Bureau claimed that the forecasts were 80% to 85% correct.

3.5.4. Ferraz (1924) submitted to an international conference a complex scheme used in Brazil. All permitted weather terms and expressions were listed. The verification rules "are based on agriculture's necessities and not on those of urban life, where people are much more severe, and unnecessarily so, in exacting absolute correctness". 24-hr. forecasts were made in three categories: weather; wind (light, moderate, etc.); temperature (stationary, small rise, moderate rise, etc.). In each category a fully verified forecast received a maximum of twelve points. Forecasts were verified at each station. Forecasts for a whole region, state, or province were considered as verified when the sum of the points obtained for n stations (well distributed) was equal to or greater than (1 + n/2)10. As an example, a forecast of "fair" received: 10 points when rain did not exceed one mm. within 30 continuous minutes, with a minimum of four hours sun; 8 points with local thunder not forecast; 8 points if unsettled without rain; 8 points with rain over one mm. within 15 minutes or with intermittent inappreciable rain within one hour; 8 points with no sunshine but clouds of good character (no dark shades); 5 points if threatening. More complicated forecasts such as "fair, changing to unsettled with rain" received points as follows: 8 points, when the first part of the forecast was verified during at least the first three hours or within the maximum period of 21 hours, and the second part, during at least the last three hours or within the maximum period of the last 21 hours, verification being strictly in accordance with the rules established (when minimum periods of three hours were considered, the limits put down by the rules could be halved); 5 points when the forecast was verified only in the maximum period; 5 points when the forecast was verified only in the second part during the last three hours, except in the case when there was no change of weather during the total period of the forecast. Forecasts made separately for day and night were verified separately, the points being halved and one added as a premium, so the maximum number of points obtainable for the 24 hours was 12. In such cases the limits established by the rules were halved.

3.5.5. While checking the accuracy of forecasts based on "perfect" prebaratics Murray and Smith (1926) devised a points system for verifying simple forecasts expressed in Beaufort letters. Table 3 contains the marking system. The results of verifying a set of 528 spot forecasts for mid-day are given in Table 4 (Table IV of Murray and Smith's paper). Precipitation forecasts were verified against the weather observed during the hour preceding mid-day. The average score for the 528 forecasts, expressed as a

/percentage

percentage, was 65. Table 4 shows that the fair weather occasions (b, bc and c) are heavily weighted by virtue of their numerical preponderance. The success in forecasting the important, but less common, phenomena, fog, hail, snow and thunder, contributes little to the final overall score. To increase deliberately the effect of the rarer categories by some artificial weighting system might not be a particularly good solution, because it is the day in and day out "rain or no rain", "sunshine or cloud" considerations which most continually affect the everyday lives of the general public, even though rarer events assume considerable importance when they do occur. These difficulties would not arise, were the forecast elements verified individually rather than in combination.

Routine use of Murray and Smith's system might encourage "hedging", since the score for a perpetual forecast of bc equals the percentage attained in practice. This consideration does not apply to the verifications of Table 4 since the checking system had not been devised when the forecasts were made.

3.5.6. Restriction of the expressions used in weather predictions to defined terms has not generally met with approval. Milham (1912) remarked that it hampered individuality and free expression of opinion even though he considered it necessary to prevent ambiguity and hedging and make possible systematic verification. In recent years many papers have dealt with the verification of numerical or simply classified forecasts but no schemes have been advanced for checking written forecasts, issued to the general public. Bleeker (1946) and Brier and Allen (1951) each refer to the undesirability of limiting the forecasters' terms. Nevertheless, there is no wish to return to subjective methods. Brier and Allen (1951) summarise the common view: "One essential for satisfactory verification is objectivity, which requires that the forecasts be explicitly stated, either numerically or categorically, thus permitting no element of judgment to enter into the comparison of the forecast with subsequent observation." And yet the forecaster who writes for the general public cannot eliminate subjective considerations; he has to interpret for the public and maintain its interest. Bleeker (1946) discusses the difficulties:

> Statements about the future value of four of five meteorological elements do not necessarily say what the weather will be like. Weather is a very complicated matter, it is namely the integral of the influences of various meteorological factors upon our senses. "With the eye", says Schmaus (1942) "we observe the colours of the sky, the clouds, the electric discharges, the precipitation. The ear tells us something of the sounds of wind and thunder, our feeling sense something about the temperature and our skin about the moisture." Every meteorologist will agree with Schmaus that it is very difficult to form a picture of the weather on a past day by combining the elements published in a year book. This can only be done in certain cases, e.g., for fine days. The reader of forecasts inevitably has the same difficulty when he tries to form an idea about the future weather from the objectively given elements of a forecast which he will experience subjectively. This subjectivity is a great handicap for the forecaster, especially if he is serving different groups of the public. Seamen, farmers, townspeople, etc. are all interested in different elements of the forecast, they apply different standards and criticize from different points of view.

How can the dilemma be resolved? The alternatives seem to be to straight-jacket the forecasts or to abandon the possibility of strict checking. Brier and Allen (1951) wash their hands of the public forecasts: "If every forecast is accompanied by a definition of the terms used, the forecaster sometimes objects that it hinders his freedom to express himself adequately. Only those forecasts which are expressed in objective terms can be verified satisfactorily as forecasts. The extent to which public forecasts satisfy the public is a different question which can be answered, it appears, only by public opinion polls; and the answer will generally contain little information about the agreement of forecast and actual weather." Bleeker, however, mentions a kind of compromise. In some meteorological services, two forecasts are compiled,

/one

one for issue to the public and one for checking: the two forecasts must not contradict each other, but the forecaster has complete liberty in the choice of his terms for the public forecasts. The system of double forecasts was introduced in Holland in 1939 for rainfall forecasts "to the great satisfaction of the acting meteorologists". Apart from allowing public forecasts to be made more colourful, such a system considerably simplifies the verification problem: more use can be made of numerical values and strict classifications, elements can be checked separately, probability statements can be made objective and difficulties of verifying area forecasts are ameliorated. The forecaster can object, sometimes with good justification, that the synopic reporting stations in a given region do not give a good objective picture of the weather over the area, although he may be able to infer from them subjectively the sort of conditions which generally obtain. It may be unfair to verify the area forecasts against the limited number of reports. If the forecaster is aware that forecasts will be checked against reports from certain stations, he may be tempted to phrase his predictions to cater particularly for those stations. This is legitimate if the forecasts are for checking purposes alone: it is not desirable when the forecasts are to be issued to the public.

3.5.7. Apart from the problems raised in the last paragraphs, certain other difficulties attend the verification of the elements of a forecast in combination. The Weather Division, H.Q., U.S.A.A.F. (1944) discuss predictions of weather type. It is remarked that if the probability of correctly forecasting one element within specified limits is 0.70, then the probability of successfully forecasting a composite of four such elements is 0.24, assuming the elements to be independent. A study of 24-hr. forecasts of air route terminal conditions is quoted. This showed that while forecasters had considerable skill in forecasting pressure, temperature and precipitation separately, they were badly beaten by persistence forecasts of the composite weather conditions. The authors conclude that if forecasts of weather type are demanded by a user, such forecasts should be furnished only in addition to, and not in place of, predictions of the respective components.

3.5.8. Most composite or semi-objective checking systems produce as their final score a "percentage correct". It was mentioned in para. 3.2.1.1. that the percentage correct depends upon the limits arbitrarily set for "correct" forecasts. Thomas (1933) for example, quotes Berlin Meteorological Office regulations which rule that an error up to two degrees in temperature is 100% correct, errors between two and four degrees are 50% correct and errors greater than four degrees are zero per cent correct. Van Everdingen (1908) gives a method used by the Royal Netherland Meteorological Institute which employs a nine-point scale for each element. According to Koppen (1893) most German institutes verified on a three-point scale, although a five-point scale was in use in Bavaria (0%, 25%, 50%, etc). And so on. Moreover, there is no agreement regarding the weights to be given to the separate elements in computing the overall percentage. Thus "percentage correct" has no meaning for those unacquainted with the details of the verification. The meteorologist himself cannot interpret the meaning of "X% correct" precisely; it cannot be inferred that the forecasts were X% correct on every occasion, neither can it be assumed that the forecasts were completely correct for X% of the occasions.

3.5.9. The foregoing paragraphs have indicated the considerable advantages to be gained in checking the meteorological content of public forecasts by having parallel forecasts made in a form suited to the verification problem. If the meteorologist is faced with written or partly written forecasts which have already been made over a long period and he is required to judge the changes in accuracy which have been effected through the years, then the parallel forecast solution is inapplicable, since there is obviously no hope of applying it in retrospect. In this case subjectivity in the comparisons is unavoidable, and the verification scheme can only attempt to minimise the inherent uncertainties. Individual elements can still be checked separately. The verifier, reading the forecasts, would place each expected state of the element into one of a number of classes. Similar treatment of

/the

the observed weather would yield a set of data which could be handled objectively from that stage, by the methods described in sections 3.2. and 3.3. The difficulties and doubts entering into the initial classification may easily be appreciated by all who are familiar with the forms in which written forecasts are given, but acceptance of a measure of subjective judgement would be implicit in a decision to check past written issues.

4. Concluding remarks

The preceding sections have shown that there is no lack of verification techniques and scores covering many of the contingencies which may arise. In most cases, the problem facing the verifier is to select the statistic or system which best suits his purpose. When fresh systems must be devised, there exists a body of experience which should enable many potential pitfalls to be avoided. When forecasting trials are held, the experiments should be designed with the verification technique and the problems of statistical significance well in mind.

5. References

The following list of references does not constitute an exhaustive bibliography, since it was not found necessary in the foregoing discussion to refer to every paper which was consulted and some papers which touch on verification indirectly may have been overlooked. An annotated bibliography has been published by Muller (1944) and a very comprehensive list of references is furnished by Bleeker (1946).

Best, W.H. and Panofsky, H.A.	1956	'Some application of statistics to meteorology', College of Mineral Industries, Pennsylvania State University.
Bijvoet, H.C. and Bleeker, W.	1951	'The value of weather forecasts', Weather, 6, p. 36.
Bilham, E.G.	1922	'A problem in economics', Nature, 109, p. 341.
Bleeker, W.	1946	'The verification of weather forecasts,' Mededeelingen en Verhandelingen, Serie A.1, No. 2.
Brier, G.W.	1944	'Verification of a forecaster's confidence and the use of reliability statements in weather forecasting', Washington Weath. Bur. Res. Pap. No. 16.
	1950	'Verification of forecasts expressed in terms of probability', Mon. Weath. Rev., 78, p. 1.
Brier, G.W. and Allen, R.A.	1951	'Verification of weather forecasts', Compendium of Meteorology, Amer.met. Soc., p. 841.
Brooks, C.E.P. and Carruthers, N.	1953	Handbook of statistical methods in meteorology, London (H.M.S.O.).
Clayton, H.H.	1869	'Verification of weather forecasts', Am. Met. J., 6, p. 211.
	1934	'Rating weather forecasts', Bull. Amer. Met. Soc., 15, p. 279.
Columbia Branch, American Met.Soc.	1952	'Panel discussion on forecast verification', Bull.Amer. Met. Soc., 33, p. 274.

/Cooke,

Cooke, W.E.	1906	'Forecasts and verifications in Western Australia,' Mon. Weath.Rev., 34, p. 23.
Crossley, A.F.	1952	'Usefulness of forecasts', Met. Mag., 81, p. 193.
	1954	'Measures of success in forecasting', Met. Mag., 83, p. 66.
Dinies, E.	1936	'Vorhersagepruefungen', Erf. Ber. deutchen Flugwetterdienstes, Sonderband 5, Teil 2.
Doolittle, H.M.	1885	'The verification of predictions', Am. Met. J., 2, p. 327.
Durst, C.S.	1954	'Variation of wind with time and distance,' Geophys. Mem., 12, No. 93.
Ekholm, N.	1907	'Uber die unperiodischen Luftdruckschrankungen und einige damit Zusammerhengende Erscheinungen, Met. Zeit., 24, p. 145.
Ferraz, de Sampaio	1924	'Verification of daily weather forecasts,' International Conf. of Directors (1923), Koninklijk Nederlandsch Met. Inst. No. 112, Appendix N.
Finley, J.P.	1884	'Tornado predictions,' Am. Met. J., 1, p. 85.
Freybe, O.	1911	'Die Traffsicherheit der Prognosen', Wetter, 28, p. 117.
Gilbert, G.K.	1884	'Finley's tornado predictions,' Am. Met. J., 1, p. 166.
Gringorten, I.I.	1951	'The verification and scoring of weather forecasts,' J. Amer. Statist. Ass., 46, p. 279.
	1955	'Tests of significance in a verification problem,' J. Met., 12, p. 179.
Gringorten, I.I., Lund, I.A. and Miller, M.A.	1955	'A program to test skill in terminal forecasting,' Air Force Surveys in Geophysics, No. 80, Geophysics Research Directorate, Cambridge, Mass.
Grossmann, L.	1898	'Die Sturme und die Sturmwarnungen an die deutche Kuste,' Aus dem Archiv der deutschen Seewarte, 21, No. 4.
Harrington, M.W.	1889	'Verifications at the Signals Office,' Am. Met. J., 6, p. 19.
Heidke, P.	1926	'Berechnung der Erfolges und der Gute der Windstarkevorhersagen im Sturmwarnungdienst, Geografika Annaler, 8, p. 310.
	1928	'Erfolg und Gute Ortlicher Vorhersagen im taglichen Wetterdienst,' Aus dem Archiv der deutschen Seewarte, 45, No. 1.

/1929

	1929	'Zur objektiven Prufung ortlicher Vorhersagen im taglichen Wetterdienst,' Beitrage Zur Physik der freien Atmosphare, 15, p. 202.
	1935	'Erfolg und Gute von Hamburger und Konigsberger Frostvorhersagen fur die nächste Nacht,' Ann. der Hydr., 63, p. 40.
Holloway, J.L. and Woodbury, M.A.	1955	'Application of information theory and discriminant function analysis to weather forecasting and forecast verification', Tech. Rep. of Met. Statistics Project of Inst. Co-op. Res., Univ. of Pennsylvania.
Johnson, D.H.	1955	'The success achieved in forecasting upper winds by orthodox and statistical techniques,' Meteorological Research Papers, Air Ministry, London, No. 953.
Jorgensen, D.L.	1956	'A graphical solution of the conventional skill score involving two categories,' Mon. Weath. Rev., 83, p. 139.
Kendall, M.G.	1948	Advanced theory of statistics, 1, (Charles Griffin and Co., London).
Köppen, W.	1893	'The best method of testing weather predictions,' Rep. of International Met. Congress, Chicago, Part 1.
Lacour, D.M.	1906	'Kritik af Stormvarselst Jenesten i Denmark i Aaret 1905, Meddelt af Meteorologisk Institut,' Arkiv for Soevaesen, p. 47.
Leight, W.G.	1953	'The use of probability statements in extended forecasting,' Mon. Weath. Rev., 81, p. 349.
Lewis, R.P.W. and McIntosh, D.H.	1952	'Some effects of coherence in meteorological time series,' Met. Mag., 81, p. 242.
Mascart, J.	1922	'La proportion des réussites dans la provision du temps,' Academie des Sciences, Comptes Rendus, 175, p. 627.
Milham, W.I.	1912	'The accuracy and verification of predictions' Meteorology, p. 404. (Macmillan, New York).
Muller, R.H.	1944	'Verification of short-range weather forecasts,' Bull. Amer. Met. Soc., 25, p. 18.
Murray, R.	1954	'On the accuracy of contour charts in forecasting upper winds,' Met. Office, London, Prof. Notes. No. 110.
Murray, R. and Smith, C.V.	1956	'The attainable accuracy of weather forecasts derived from 24-hr. prebaratic charts.' Unpublished.

/Peirce,

Peirce, C.S.	1884	'The numerical measure of the success of predictions,' Science, 4, p. 453.
Priestley, C.H.B.	1945	'Accuracy of upper air forecasts,' Meteorological Research Papers, Air Ministry, London, No. 222.
Sawyer, J.S.	1953	'The correlation coefficient, as a measure of success in forecasting,' M.O.21, Tech. Notes, No. 76, Met. Office, Dunstable.
Schmauss, A.	1942	'Das Problem der Wettervorhersage,' Leipzig.
Thomas, H.	1933	'Zur Prufung der synoptischen Vorhersagen,' Met. Zeit, 50, p. 29.
Thompson, J.C. and Brier, G.W.	1955	'The economic utility of weather forecasts,' Mon. Weath. Rev, 83, p. 249.
Vernon, E.M.	1953	'A new concept of skill score for rating quantitative forecasts,' Mon. Weath. Rev., 81, p. 326.
Wagemann, H.	1935	Prognosenprufung, Annalen der Hydrographie, 63, p. 270.
Wallén, A.	1921	'Sur le Contrôle des Annonces de tempêtes,' Geograf. Ann., 3, p. 267.
Weather Division, H.Q., U.S.A.A.F.	1944	'Critique of verification of weather forecasts,' Weather Division, H.Q. U.S.A.A.F., Rep. No. 669.
Weather Information Branch, H.Q., U.S.A.A.F.	1943	'Short range forecast verification programme,' Weather Information Branch, H.Q., U.S.A.A.F., Rep. No. 602.
	1944	Ibid., Supplement to Rep. No. 602.
Yule, G.U. and Kendall, M.G.	1937	Introduction to the theory of statistics, (Charles Griffin and Co., London).

Table 1. Contingency tables of precipitation forecasts by A and B (after Vernon 1953)

Forecasts by A

Observed	0	0.01-0.20	0.21-0.50	0.51-1.00	⩾ 1.01	Total
0	65	5	2	1	1	74
0.01 - 0.20	9	6	3	3	1	22
0.21 - 0.50	4	5	4	2	1	16
0.51 - 1.00	1	2	3	3	1	10
⩾1.01	1	1	2	2	2	8
Total	80	19	14	11	6	130

Conventional skill score = 0.362
Deviation " " = 0.498
Quadratic " " = 0.602

Forecasts by B

Observed	0	0.01-0.20	0.21-0.50	0.51-1.00	⩾ 1.01	Total
0	65	9				74
0.01 - 0.20	10	6	6			22
0.21 - 0.50	5	4	4	2	1	16
0.51 - 1.00			4	3	3	10
⩾1.01				6	2	8
Total	80	19	14	11	6	130

Conventional skill score = 0.362
Deviation " " = 0.639
Quadratic " " = 0.824

G. 23021/IKW/7/57/50

Table 2. Contingency tables of biased precipitation forecasts by A and B (after Vernon 1953)

Forecasts by A

Observed	0	0.01-0.20	0.21-0.50	0.51-1.00	⩾ 1.01	Total
0	65		9			74
0.01 - 0.20	9		13			22
0.21 - 0.50	4		12			16
0.51 - 1.00	1		9			10
⩾ 1.01	1		7			8
Total	80		50			130

Conventional skill score = 0.323
Deviation " " = 0.469
Quadratic " " = 0.582

Forecasts by B

Observed	0	0.01-0.20	0.21-0.50	0.51-1.00	⩾ 1.01	Total
0	65		9			74
0.01 - 0.20	10		12			22
0.21 - 0.50	5		11			16
0.51 - 1.00			10			10
⩾ 1.01			8			8
Total	80		50			130

Conventional skill score = 0.310
Deviation " " = 0.482
Quadratic " " = 0.631

G.23021/IKW/7/57/50

Table 3. Scoring system for weather forecasts (Murray and Smith 1956)

Observed Weather

Forecast Weather	b	bc	c	f	pr	prs	ps	ph	tl	r	rs	s
b	5	4	3	3	1	1	1	1	0	0	0	0
bc	4	5	4	3	2	2	2	2	0	0	0	0
c	3	4	5	3	2	2	2	2	0	0	0	0
f	3	3	3	5	0	0	0	0	0	0	0	0
pr	1	2	2	0	5	4	3	4	3	3	2	1
prs	1	2	2	0	4	5	4	4	3	2	3	2
ps	1	2	2	0	3	4	5	4	3	1	2	3
ph	1	2	2	0	4	4	4	5	3	2	2	2
tl	0	0	0	0	3	3	3	3	5	3	3	3
r	0	0	0	0	3	2	1	2	3	5	4	3
rs	0	0	0	0	2	3	2	2	3	4	5	4
s	0	0	0	0	1	2	3	2	3	3	4	5

Table 4. Results of Press forecasts (Murray and Smith 1956)

Observed Weather

Forecast Weather	b	bc	c	f	pr	prs	ps	ph	tl	r	rs	s	Total
b	18	7	11	3	1					1			41
bc	30	70	60	7	22		1			22			212
c	5	32	44	1	4		1	1		27		1	116
f			5							2			7
pr	3	13	14		18			2		14	1		65
prs		1						1					2
ps						1	2	1					4
ph							1						1
tl													0
r	3	16	23	1	8		2			25	1		79
rs			1										1
s													0
Total	59	139	158	12	53	1	7	5	0	91	2	1	528

Forecast verification: Appendix I

Some representative values of the verification indexes

1. A series of 212 routine 24-hr. temperature forecasts for October and November 1956 have been checked to compare values of some of the verification indexes.

2. Tables A and B are contingency tables for routine forecasts versus the observed values and for persistence forecasts versus the observed values. Measures of postagreement ($X_{ii}/(A_i)$) and prefigurance ($X_{ii}/(B_i)$) (para. 3.3.1.1. of the survey) suggest that the forecasts had greater success over the middle of the range than at the extremes. The routine forecasts were more successful than the persistence forecasts in most categories, but there is a close relation in the values achieved by the two techniques.

3. The tetrachoric tables C and D classify the forecasts dichotomously according as the temperatures forecast were below the seasonal normal or not. The routine forecasts improved upon persistence for both prefigurance (c/o) and postagreement (c/p) (para. 3.2.2. of the survey), but the gain was greatest for postagreement.

4. Table E contains values of some of the more useful of the indexes discussed in the survey. The numerical values of these indexes vary widely but it is noteworthy that they all place the two sets of forecasts in the same order of merit.

5. Table F gives values of a few of the indexes, based on the same data, for which the forecast changes rather than the forecasts themselves are considered. These indexes are generally smaller than their counterparts in Table E, since the success they represent owes nothing to persistence.

Table A. Contingency table for routine 24-hr. forecast temperatures versus observed values

		Forecast temperature range in degrees Fahrenheit									Totals	Prefigurance
		36-38	39-41	42-44	45-47	48-50	51-53	54-56	57-59	60-62		$(X_{ii}/(B_i))$
Observed temperature range in degrees Fahrenheit	36-38	1	2								3	.33
	39-41	2	1	3	3						9	.11
	42-44	1	1	1	4	3					10	.10
	45-47			1	13	10	3				27	.48
	48-50			2	12	28	7	3			52	.54
	51-53		1		2	20	13	8	3		47	.28
	54-56					4	8	14	4		30	.47
	57-59						2	8	10	7	27	.37
	60-62						1		6		7	.00
	Totals	4	5	7	34	65	34	33	23	7	212	
Postagreement $(X_{ii}/(A_i))$		.25	.20	.14	.38	.43	.38	.42	.43	.00		

G. 23021/IKW/7/57/50

Table B. Contingency table for persistence 24-hr. forecast temperatures versus observed values

		Forecast temperature range in degrees Fahrenheit									Totals	Prefigurance
		36-38	39-41	42-44	45-47	48-50	51-53	54-56	57-59	60-62		$(X_{ii}/(B_i))$
Observed temperature range in degrees Fahrenheit	36-38	1	2								3	.33
	39-41	1	1	2	2	3					9	.11
	42-44		2		5	2		1			10	.00
	45-47		2	3	5	8	7	1	1		27	.19
	48-50	1		4	9	23	13	1	1		52	.44
	51-53			2	4	14	12	13	2		47	.26
	54-56				1	2	8	11	6	2	30	.37
	57-59						4	7	11	5	27	.41
	60-62							1	5	1	7	.14
	Totals	3	7	11	26	52	44	35	26	8		
Postagreement $(X_{ii}/(A_i))$		.33	.14	.00	.19	.44	.27	.31	.42	.13		

Table C. Tetrachoric table classifying routine forecast and observed temperatures as below or not below normal

		Routine Forecast		
		not below	below	total
Observed Temperature	not below	57	26	83
	below	21	108	129
	total	78	134	212

Prefigurance

$\frac{c}{o}$ = .69

Postagreement

$\frac{c}{p}$ = .73

Table D. Tetrachoric table classifying persistence forecasts and observed temperatures as below or not below normal

		Persistence Forecast		
		not below	below	total
Observed Temperature	not below	53	30	83
	below	32	97	129
	total	85	127	212

Prefigurance

$\frac{c}{o}$ = .64

Postagreement

$\frac{c}{p}$ = .62

Table E. Verification indexes for 24-hr. temperature forecasts (forecast versus observed values)

Classification	Index	Routine Forecast	Persistence
dichotomous (below normal or not)	Peirce	.52	.39
	Clayton	.54	.39
	Yule's V	.53	.39
	skill score	.53	.39
manifold	skill score	.25	.17
	Vernon's S_d	.60	.50
	Vernon's S_q	.81	.72
	Pearson's C	.79	.73
	Pearson's C (adjusted)	.84	.78
	Tchuprow's T	.46	.38
	correlation (grouped)	.81	.72
continuous distribution (forecast v. observed value)	mean error	+.17°F	-.26°F
	root mean square error	3.06	3.63
	correlation coefficient	.82	.74

Table F. Verification indexes for 24-hr. temperature forecasts (forecast versus observed changes)

Classification	Index	Routine Forecast
dichotomous (colder or not colder)	Peirce	.31
	Clayton	.31
	Yule's V	.31
	skill score	.31
Continuous distribution (forecast change v. observed change)	mean error	+.17°F
	root mean square error	3.06
	correlation coefficient	.61

Forecast verification: Appendix II

Tests of statistical significance in forecast verification

1. Representativeness of data

1.1. When forecasts are verified it is necessary to estimate whether the results of the verifications are likely to be of general validity. This question is particularly relevant when the purpose of the checking is to decide the relative merits of two sets of forecasts. The use of tests of statistical significance in this connexion has increased in recent years, although it is still far from general. Owing to the underlying assumptions, standard treatments given in statistical text-books are often inadequate to deal with the meteorological case, the most frequently invalid supposition being that successive data are independent.

1.2. The representativeness of the weather types occurring during the period of the verifications must also be considered. Suppose, for example, that a forecast trial showed one forecast technique to achieve a (statistically) significantly greater success than another. If the weather during the trial happened to be predominantly anticyclonic, then the significance of the difference in the forecast successes would have been established for anticyclonic conditions, but the significance tests would have nothing to say about the success achieved in a generality of weather types. In brief, verification samples are rarely random.

2. Tests on independent data

2.1. If the data can be selected so that successive forecasts and successive occurrences are independent, then statistical significance can be estimated by standard methods.

2.2. For dichotomous and manifold classifications questions of significance are best resolved by applying a chi-squared test to the basic contingency table, although this test should be used only when the expected frequency is not less than five in any cell (Brooks and Carruthers, 1953).

In some cases formulae for the standard errors of verification indexes are available. For Yule's V, in the notation of section 3.2. of the survey, the standard error is given by the square root of the expression,

$$\frac{1}{S}\left[1 - V^2 + \frac{(V + \frac{1}{2}V^3)(2o-s)(2p-s)}{(op(s-o)(s-o))^{\frac{1}{2}}} - \frac{3V^2s^2\,(p(s-p) - o(s-o))}{4op\,(s-p)(s-o)}\right]$$

Yule (1912) derived this expression assuming that the observed cell frequencies may be used in place of the population cell frequencies and that successive trials are independent.

2.3. The standard errors of many of the statistics discussed in section 3.4. of the survey (continuous distribution) are well-known.

For the mean error, or bias, the standard error is given by $\sigma/\sqrt{N}$ where σ is the standard deviation of the N forecast errors.

The standard error of the difference between two biases is $\sqrt{(\sigma_x^2+\sigma_y^2)}/\sqrt{N}$ where σ_x and σ_y are the standard deviations of the two error distributions. For small (uncorrelated) samples Student's 't' test is available.

/The

The standard error of the difference between two standard deviations is given by

$$\left[\text{standard error of } (\sigma_1 - \sigma_2)\right]^2 = \frac{\mu_4 - \mu_2^2}{4\mu_2 N} + \frac{\nu_4 - \nu_2^2}{4\nu_2 N},$$

where μ_4, μ_2 and ν_4, ν_2 are the fourth and second moments of the forecast error distribution and N is the number of cases (Yule and Kendall, 1937). When the errors are normally distributed, this formula reduces to $(\sigma_1^2 + \sigma_2^2)/2N$

Correlation coefficients are usually tested for significance by means of Fisher's Z-test. Kendall (1948) advises against the use of the standard error in this case since the usual assumptions of normality of the sampling distribution, etc. do not hold unless N is very large; how large N should be it is not easy to say but values as high as 500 are unsatisfactory.

Vector errors require special consideration. Components may be treated individually but it is often convenient to sum departures from the vector mean error over all directions to get a scalar variate which can be handled along the lines discussed above.

3. Serially correlated data

3.1. An attempt can be made to avoid serial correlation by spacing the forecasts so that the persistence effects are minimised, but, as Gringorten (1955) rightly remarks, it is difficult to avoid the problem completely because there is persistence of weather patterns on all scales, even from month to month. Gringorten continues: it is better to live with this difficulty and assume that it will be negligible. This view is hardly tenable; the effects of serial correlation are often far from negligible; for example, even when the serial correlation in a set of observations is as low as 0.35, for the purpose of estimating the accuracy of the mean, the equivalent number of independent values is only one half of the total number of observations.

3.2. Although the difficulties arising in extending the theory of significance tests to cover cases where successive values are not independent have not been generally solved, a rigorous approach is possible when the verification statistic can be expressed as a mean value.

For a series of N terms, having standard deviation σ and serial correlations p_1 between consecutive terms, p_2 between alternate terms, etc., the standard error of the mean value is

$$\sigma\left[1 + \frac{2}{N}\left\{(N-1)p_1 + (N-2)p_2 + \ldots\ldots + p_{N-1}\right\}\right]^{\frac{1}{2}} \Big/ N^{\frac{1}{2}}.$$

When the mean values of two such series, $x_1, x_2 \ldots\ldots x_N$, $y_1, y_2 \ldots\ldots y_N$, are to be compared, the square of the standard error of the difference between the means is given by

$$\frac{\sigma_x^2 + \sigma_y^2}{N}\left[1 + \frac{2\left\{P\sigma_x^2 + Q\sigma_y^2 - \sigma_x\sigma_y(r + 2T)\right\}}{\sigma_x^2 + \sigma_y^2}\right]$$

/where

where, $P = \frac{1}{N} \left((N-1)p_1 + (N-2)p_2 + \ldots\ldots\ldots\ldots\ldots\ldots + p_{N-1} \right)$;

$Q = \frac{1}{N} \left((N-1)q_1 + (N-2)q_2 + \ldots\ldots\ldots\ldots\ldots\ldots + q_{N-1} \right)$;

$T = \frac{1}{N} \left((N-1)t_{1,2} + (N-2)t_{1,3} + \ldots\ldots\ldots\ldots\ldots\ldots + t_{1,N} \right)$;

r is the correlation between the x and y; p_1, p_2, etc. are correlations in the x; q_1, q_2, etc. are correlations in the y; $t_{i,j}$ are cross-correlations between the x_i and y_j (Johnson, 1955). These formulae can be applied when N is sufficiently large for the following assumptions to hold:

(i) the sample standard deviations and correlations may be used as estimates of the population values;

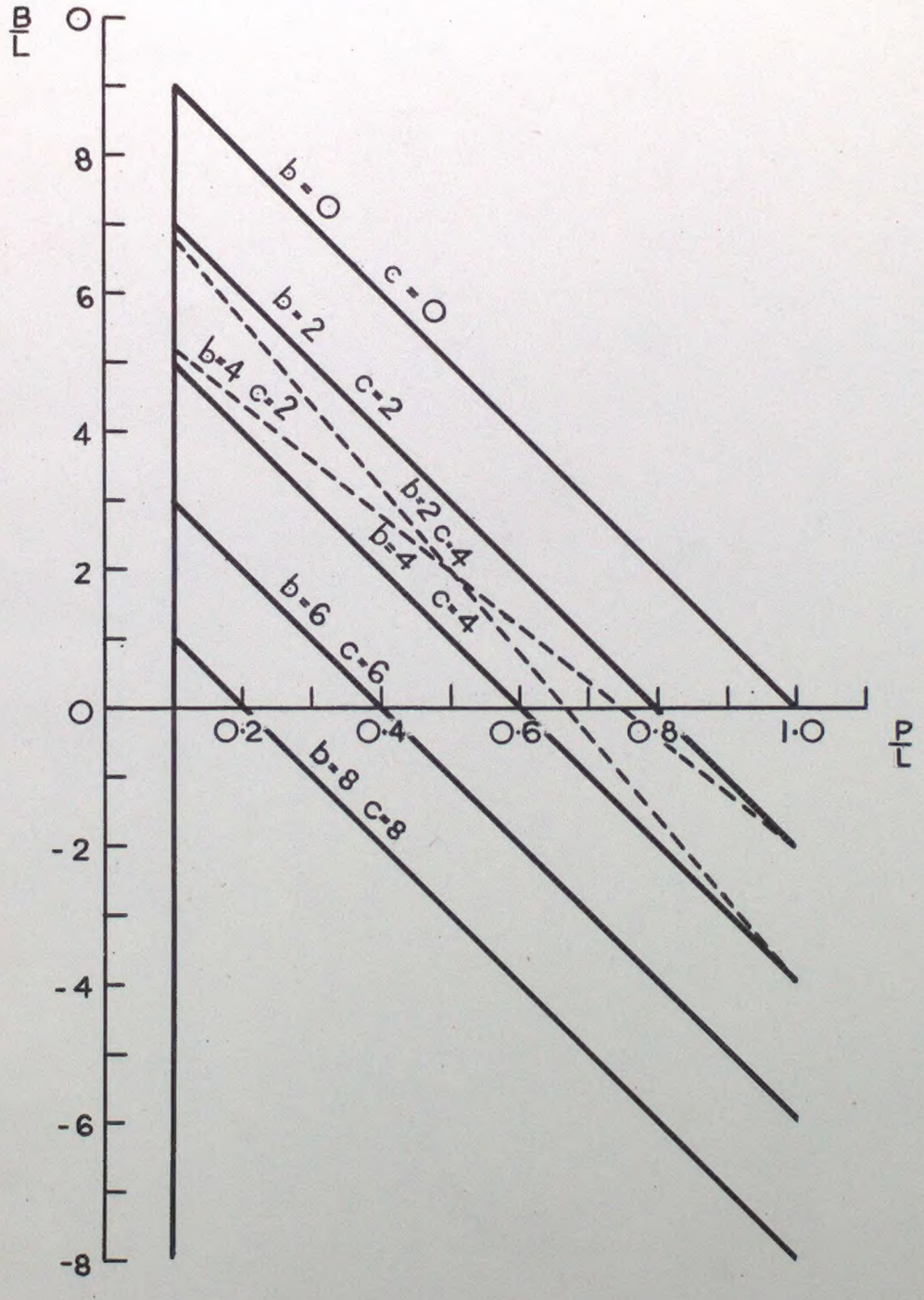
(ii) the sampling distribution of the bias is approximately normal.

These conditions become better fulfilled as N increases. When the serial correlations are about .3, a good approximation is probably obtained for $N > 60$ but when higher serial correlations exist a larger number is required. In estimating this number it may be helpful to calculate the "equivalent number of independent observations" (Lewis and McIntosh, 1952).

3.3. The above method of estimating the standard error can be applied to the bias (continuous distribution, para. 3.4.1.), when the x and y are simply the forecast errors. The serial correlations in forecast errors are usually substantially smaller than the serial correlations in the forecast or observed values, but it must not be assumed that they are negligible. In a set of 18-hr. upper wind forecasts, for example, significant serial correlations in the forecast errors of about 0.2 were found. The errors made in upper wind forecasts issued from two independent stations were correlated with a coefficient as high as 0.6 (Johnson, 1955).

Other indexes which essentially award a mark for each forecast made can be adapted for similar treatment. Examples are: percentage correct (para. 3.2.3.); Vernon's scores (para. 3.3.1.4.); Gringorten's score (para. 3.3.1.9.); Brier's Index (para. 3.3.2.2.); variance of the forecast errors (para. 3.4.2.).

3.4. The problem of allowing for serial correlation in dealing with other types of verification score, such as the correlation coefficient, has not yet been satisfactorily solved, to the writer's knowledge. A possible approach is to estimate an "equivalent number of independent terms" (Lewis and McIntosh, 1952) or an "equivalent number of independent comparisons" (Johnson, 1955), but this method involves the working assumption that the equivalent number of independent terms is about the same for the correlation coefficient as for a mean value.

Fig. 1

The relation between the specific industry factor, (P/L) and the economic benefit factor, (B/L) for 10 unfavourable days out of 100, total loss being smaller than the total cost of preventive measures (After Bijvoet and Bleeker)